# The Usability of Speech and/or Gestures in Multi-Modal Interface Systems

Farzana Alibay [*]
Department of Computing,
Macquarie University
2019, Sydney
NSW, Australia
farzana.alibay@gmail.com

Manolya Kavakli [†]
Department of Computing,
Macquarie University
2019, Sydney
NSW, Australia
manolya.kavakli@mq.edu.au

Jean-Rémy Chardonnet [‡]
Le2i FRE2005, CNRS, Arts et Métiers, Univ. Bourgogne Franche-Comté, Institut Image
71100, Chalon-sur-Saône
France
jean-remy.chardonnet@ensam.eu

Muhammad Zeeshan Baig [§]
Department of Computing,
Macquarie University
2019, Sydney
NSW, Australia
baig.mzeeshan@gmail.com

## ABSTRACT

Multi-Modal Interface Systems (MMIS) have proliferated in the last few decades, since they provide a direct interface for both Human Computer Interaction (HCI) and face-to-face communication. Our aim is to provide users without any prior 3D modelling experience, with a multi-modal interface to create a 3D object. The system also incorporates help throughout the drawing process and identifies simple words and gestures to accomplish a range of (simple to complex) modeling tasks. We have developed a multi-modal interface that allows users to design objects in 3D, using AutoCAD commands as well as speech and gesture. We have used a microphone to collect speech input and a Leap Motion sensor to collect gesture input in real time. Two sets of experiments were conducted to investigate the usability of the system and evaluate the system performance using Leap Motion versus keyboard and mouse. Our results indicate that performing a task using speech is perceived exhausting, when there is no shared vocabulary between man and machine, and the usability of traditional input devices supersedes the usability of speech and gestures. Only a small ratio of participants, less than 7% in our experiments were able to carry out the tasks with appropriate precision.

## CCS Concepts

•**Human-centered computing → Human computer interaction (HCI); Interaction design; User interface design;**

## Keywords

Gesture; Speech; Semantics; Emotion Recognition; Kinect; 3D object; Leap Motion

## 1. INTRODUCTION

In the last decades, many efforts have been made to improve the performance of uni-modal and multi-modal interpreters. A Multi-Modal Interface System (MMIS) aggregates two or more user input modes in an interconnected fashion with multimedia output. The user input can be speech, pen, touch, manual gesture, gaze, head, and body movements [8]. One of the main problems in the field of MMIS is to develop systems that utilize human behavior and language to interact with computers. Speech input has been extensively used in smartphones especially for developing commercial products. Another popular input mode is gestures, inspiring many researchers to develop gesture recognition systems and algorithms for HCI with practical applications [3]. There is some evidence suggesting that a MMIS not only improves handling and reliability of the system, but also task completion rates compared to unimodal systems [7]. However, the need for a multi-modal interface instead of a single input interface is relatively less explored. In this paper we investigate the usability of speech and hand gestures in a multi-modal versus traditional interface.

A typical MMIS design consists of a recognition system that translates human tasks into recognizable computer signals. Once the human input has been identified, the next step is to interpret the

[*]Farzana Alibay was a MRes exchange student at Macquarie University from Institut Image, France.

[†]Manolya Kavakli is an Associate professor at Department of Computing at Macquarie University, Australia.

[‡]Jean-Rémy Chardonnet is an Assistant professor at Institut Image, France.

[§]Muhammad Zeeshan Baig is a PhD Scholar at Department of Computing, Macquarie University, Australia.

inputs and aggregate them to achieve a desired output. Most examples in the literature use speech and pen input in MMIS design [5]. An early example was from Bolt's "Put That There" multimodal system that combined speech and pointing gestures to move an object [2].

Some recent applications have also utilized gesture input and combined it with speech to draw sketches and compare them with hand-drawn sketches [4]. Most of these systems have used Kinect and Leap Motion to recognize gesture input. The speech has provided an extra dimension for information required to interact with the computer in cases such as coloring or rotating the object [9].

While combining two input sets is beneficial for some applications, it may not be so beneficial or preferable in some others. For example, in a modeling software, many complex words are used to draw a 3D object. The users have to be familiar with the vocabulary and have to learn how to navigate in the 3D space. This research investigates: the effectiveness of using simple words and gestures to design or navigate in the 3D space, the combination of speech and gesture inputs to perform design tasks and facilitate the design process, the easiness for a user to use speech and gesture recognition systems instead of a keyboard and mouse, and the ideal type of communication channel for designer-computer interaction.

The research aims to provide users without any prior 3D modelling experience, with a multi-modal interface to create a 3D object. The system also incorporates help throughout the drawing process and identifies simple words and gestures to accomplish a range of (simple to complex) modeling tasks.

In Section 2, we explain the methodology of the project. Section 3 sheds light on the experimental process. Evaluation of the experimental results is given in Section 4, before conclusion.

# 2. RESEARCH METHODOLOGY

To test the usability of multi-modal and uni-modal input systems, we have developed a system to model a 3D object using gesture and speech inputs. In this section, we will describe the system design and architecture. As the design concept, we developed a model to convert speech and gesture actions into commands given in AutoCAD.

## 2.1 System Specifications

We have used an Intel Core i7 desktop PC, with a Microsoft Windows 8.1 operating system. For gesture recognition, we have used a Leap Motion sensor, instead of Kinect, since our pilot experiments showed that Kinect 1.0 does not allow recognition of users' fingers [10]. Therefore, Leap Motion and its API have been chosen for gesture input, since it offers facilities for finger recognition [6]. For speech recognition and synthesis, we have used a typical microphone and the Microsoft Speech Recognition API. We have chosen to use the AutoCAD 2017 3D modeling software for the users to design an object. To create an AutoCAD plugin, ObjectARX 2017 SDK was installed [1]. For implementation of the system, we used C# with the Microsoft Visual Studio 2015 Environment.

## 2.2 Design Concept

For experimental purposes, we have identified a classical chair example to draw and manipulate using multi-modal input. We have analyzed the necessary processes for this design concept.

### *2.2.1 Manipulation and Object Identification*

The classical manipulation processes to draw an object involve functions such as select, move, rotate, delete, copy, and scale. We have defined possible actions using speech and gesture inputs to apply the above-mentioned manipulation functions. For example, to rotate an object, the user has to first select it. The object can be selected using speech or gesture. To select it using gestures, the user needs to navigate the cursor to it and perform a clicking gesture. To perform the same actions using speech and gestures, first the user needs to navigate the cursor to the object and then say the keyword "select" to select the object. To rotate it with a gesture, the user needs to hang on to the clicking gesture and rotate it with the hand position. If the user wants to perform rotation with speech, the keyword "rotate to" followed by the direction of rotation, which should be 90, 45 or 180 degrees, is used. The same set of AutoCAD commands has been used for all other manipulation actions: first select the object and then use keywords to manipulate it. In summary, to perform a task using both gestures and speech, the process is much more complicated.

### *2.2.2 Drawing identification and manipulation*

To draw a chair, first we need to draw shapes in AutoCAD such as a rectangle, a cylinder, an arc, etc. We also need to have the ability to round the shapes and give some height and thickness to a surface. Finally, we also incorporate the functionality of applying a texture, material and color. For example, if we need to draw a box using speech, first we need to say "I want to draw a box"; the system will look for the word "Draw" in the speech and find the shape, which in this case is a box. After the object has been selected, the next step is to specify the position, which can be defined using the command "the position is x, y, z", where x, y, z are coordinates in 3D. The third step is to give the object a size or dimensions; to achieve this task the user needs to say "the size is x, y", where x and y are the length and the width. We also have to mention the height of the object by saying "the height is z", where z is the height of the specified object. For a circular object, the user needs to mention the radius instead of the size.

If the user wants to perform the same tasks using gestures, all he needs is to use the hands to locate the position and click on the specific icon to draw a shape, using the clicking gesture; then, with the help of the click and hold function, the size and height of the object should be adjusted. To assign a color or material to the object, after selecting the object, a speech command "the material or color is" can be used. In this project, only wooden material and gray color can be assigned to an object.

Usually, in a modeling software, there are two possible ways to manipulate the camera view; either using a mouse or the orbit. The orbit is the easiest way to move the camera by clicking directly on the cube (top, right, left, back, down, front or the corner right/back or right/front). With speech, we can move the camera using classical directions such as 'move camera vertically and horizontally' and 'zoom in and out'. We also orientate the camera by specifying the number of degrees and the direction, stating "orientate the camera to 45 degrees on the right". Using speech, if no number is specified with the direction, then the default value is applied (1 degree, or 1 cm). Using gestures, the camera can be activated or deactivated: when the system detects a closed hand followed by an open hand, it activates or deactivates the camera.

Table 1 shows a detailed description of the words used in speech and the corresponding AutoCAD commands. For gestures, the user needs to utilize the clicking gesture on the icon or a tool to enable that command.

During the experiment, we have implemented a user-assistance system, so that the user enables the assistance by saying "Help me, please". This instantiates the help sequence and offers a way to perform an action. The system also provides assistance, while performing an action. For example, if the user chooses to draw a box, the system will ask him to choose the position. Once the user

**Table 1: Speech and corresponding AutoCAD commands**

| Speech | AutoCAD commands |
|---|---|
| Box, rectangle, square, bars, layer | BOX |
| Cylinder, tube | CYLINDER |
| Cone | CONE |
| Wedge | WEDGE |
| Sphere | SPHERE |
| Torus, donut | TORUS |
| Arc | ARC 3 points |
| Extrude | EXTRUDE |
| Fillet, FilletEdge, round | FILLETEDGE |
| Thicken | THICKEN |
| Move, Displacement | MOVE |
| Copy, duplicate, clone | COPY |
| Remove, Delete | DELETE |
| Scale | SCALE |
| Rotation, rotate | ROTATE |
| Undo | UNDO |
| Finish | ENTER (to finish an action) |
| Select all | SELECTALL |
| Select last | SELECTLAST |

chooses the position either by speech or gesture, the system asks him to choose the size and height. The system anticipates the next step for the current action and guides the user to perform it.

## 3. IMPLEMENTATION

To implement the design concept, we have created a dll plugin for AutoCAD containing 4 main classes: MySpeech, LeapListener, MyMain, and MyDrawAutocad. The MySpeech class starts the speech recognition function and sends an event when a speech is recognized. LeapListener first initializes the gesture recognizer and defines all the gestures to be recognized and sends an event when a gesture is recognized. MyMain receives the speech and gesture events, interprets them and sends the right command to AutoCAD. MyDrawAutoCad contains all the functions to draw or manipulate the object or the camera. Figure 1 shows the implemented MMIS structure.

### 3.1 AutoCAD Commands

In AutoCAD, a user can use speech and gesture inputs to execute a command. For speech, the user needs to say the desired command such as "I want to draw a box". With gestures, the user is required to navigate the cursor towards the desired icon and do the clicking gesture. To load the plugin in AutoCAD, the "netload" functionality of AutoCAD has been utilized. Almost all the main functions of AutoCAD have been used such as shapes, color, material, copy, rotate, delete and camera manipulation, etc.

### 3.2 Speech recognition

The speech recognition block starts with the initialization of the Microsoft speech recognition and synthesizer API. The next step is to generate grammar for the speech recognition engine. For grammar building, almost all the main functionalities of AutoCAD has been incorporated in the grammar along with a simple sentence structure. After successful initialization of the speech recognition API and detection of the audio input device, the speech recognition process starts. When a speech is detected with a reasonable confidence level, the event is sent to the main block for further interpretation. Figure 2 shows a block structure of the speech recognition process with the corresponding events.

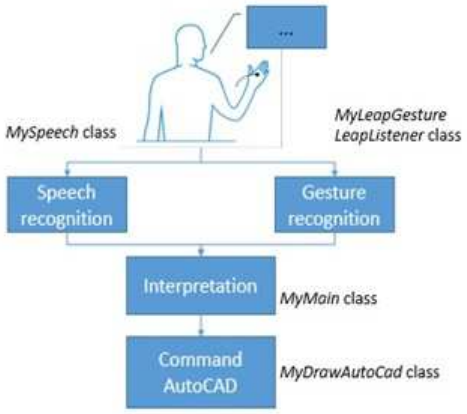



**Figure 1: MMIS structure of the implemented system**

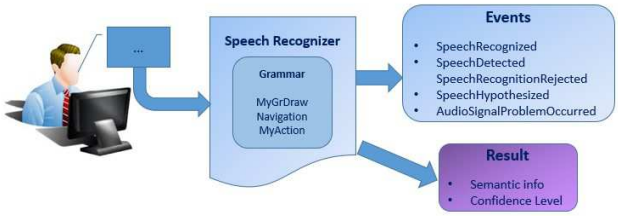



**Figure 2: Speech recognition module block diagram**

The speech recognition module has some limitations that have to be considered to effectively translate speech into text, such as noise, syllable lengths, and clarity of speech by the user.

### 3.3 Gesture recognition

In this project, the right hand has been used to control the mouse cursor: when the right hand is closed and then opened, it simulates the click of the left mouse button. The left hand has been used to manage the camera view. An event will be sent when the system recognizes that the hand is moving horizontally or vertically or the hand rotates or the hand is closed or open. The opening and closing of the left hand have been used to activate or deactivate the movement of the camera. To recognize gestures, the Leap Motion sensor and its SDK have been used. The Leap Motion sensor has the ability to recognize gestures, identify hands, the number of fingers, the grasping strength, the velocity and direction of movements along with yaw, roll and pitch. The velocity and movement are manipulated to control the cursor in the X and Y directions. The main block receives the speech and gesture events and interprets them to apply the correct AutoCAD commands.

## 4. EVALUATION

We evaluated the usability of the system using both quantitative and qualitative assessments. Every user performed two experiments, first using the keyboard and mouse inputs, and second using speech and gesture inputs. After the experiment, user feedbacks were collected through a questionnaire. This questionnaire includes questions to measure the perceived user performance, fa-

tigue and cognitive load. The time required for completing the whole experiment is 60-90 minutes per user depending on how familiar the user is with AutoCAD.

## 4.1 Description of the experimentation

The first experiment is to draw a chair using the keyboard and mouse in AutoCAD. For participants who did not know AutoCAD, a step-by-step guide was provided to draw the chair. They had to manipulate a camera view to be able to draw the chair correctly.

For the second experiment, the participants had to get familiar with how to manage both hands: the right hand to simulate the mouse and the left one to control the camera view. Then, they could start to draw the chair. For this experiment, written speech or gesture instructions were also provided.

## 4.2 Testing the application

To evaluate the system performance and find out if it is easier to draw a chair using speech and gestures than using a keyboard and mouse, a log has been created to store the experimental data, which contains the history of commands. Thus, we are able to extract the following information:

- Number of detected speech commands
- Number of recognized speech commands
- Number of low-confidence speech recognition
- Number of hypothesized speech commands
- Number of rejected speech commands
- Number of audio signal issues

To evaluate the system, eight individuals (6 men and 2 women) were selected for the experimentation and none had prior knowledge of AutoCAD. All of the participants were computer science students and academics between 20 and 50 years old. All were foreigners but spoke English fluently. At the end of the experiment, the participants were expected to fill in a questionnaire to assess the qualitative performance of the system. In the questionnaire, each question is asked twice, in order to compare their status when performing the task with the keyboard and mouse and with gesture and speech. There are a number of questions related to the performance of the commands, fatigue felt and the participants' perception in interacting with a computer.

## 4.3 Log file evaluation

A quantitative analysis has been performed using the data recorded in the log file of each set. Half of the users completed the task in 30 minutes and half took 45 minutes. No user finished drawing the chair by gesture and speech, since they gave up. Those who completed in 45 minutes drew with high precision and the other group drew the chair without obeying the guidelines.

With the log file, we were able to extract information about speech recognition. Figure 3 illustrates different audio signal issues identified during the experiment. 88% of the audio signal issues originated from the signal being too soft - meaning that a soft voice caused too much attenuation on the signal. 9% of the audio signal issues originated from the signal being too noisy. The rest of the audio signal issues originated from being either too loud, too slow or too fast. Figure 4 shows the comparison between recognized, rejected and hypothesized words. Almost 77% of words were hypothesized (detected with low certainty), 14% were rejected and only 7% of words were accepted. The main problem in speech

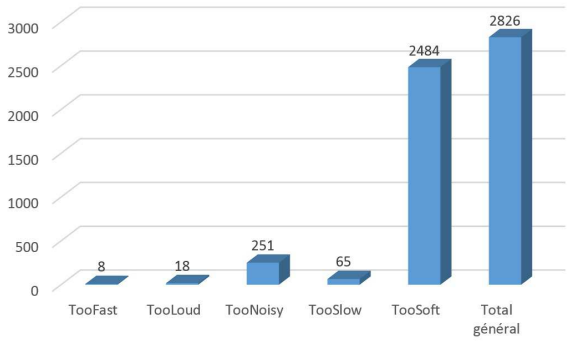



**Figure 3: Audio signal issues during speech recognition**

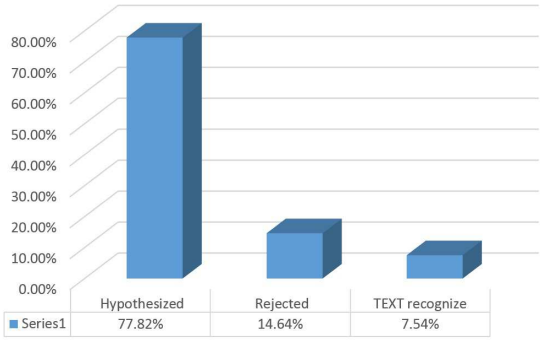
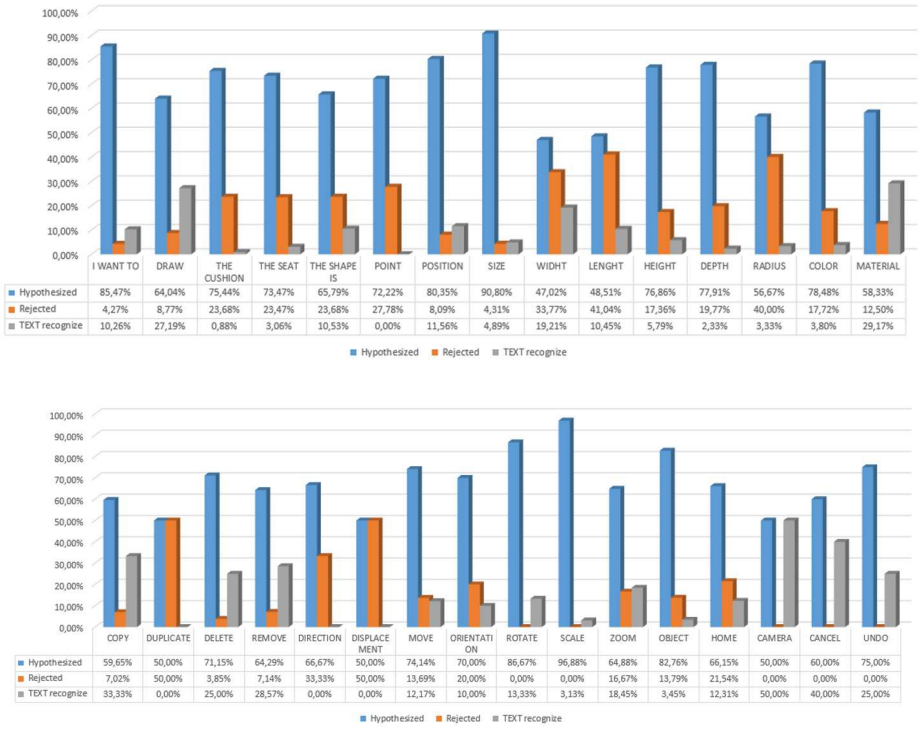



**Figure 4: Percentage of hypothesized, rejected and recognized words**

recognition was the hesitation in participants' voice, while speaking. The system expected a complete sentence but the participant expressed only the portion of the sentence such as a number to define the size. Therefore, the system misunderstood the words and even recognized the word "zoom", when the user did not say anything. Figure 5 shows the percentage of sentences recognized and rejected for drawing or manipulating the object. In general, the speech signals were highly hypothesized. It has been noticed that the participants found hard to use speech recognition. The system recognized the words "Draw" (27%), "Copy" (33%) and "Material"(29%) rather well, but "Size", "Height", "Depth", "Scale", "Color", "Radius" have less than 10% recognition rate. It was not possible to exploit the data collected in the log file for gesture recognition. However, we analysed the video records to identify how comfortable the users were through observation. We found that almost none was comfortable with camera manipulation; sometimes the system recognized the hand closed as hand open, even if the hand was partially closed. On the contrary, the system did not recognize the hand closed for the clicking gesture. It was hard for the user to specify the size by using gestures and for left-handed users, it was very difficult to control the mouse.

Through the questionnaires, we concluded that, in general, it was not easy to draw the chair and manipulate the camera in AutoCAD, but it was easier to perform these actions using a keyboard and mouse rather than gesture and speech. The users felt more exhausted while using gesture and speech than using a keyboard and mouse. For the users, it was more natural to use a keyboard and mouse than speech and gesture. They were more satisfied by the response of the computer and more engaged. They felt more frustrated by gesture and speech inputs but appreciated the help and




| | I WANT TO | DRAW | THE CUSHION | THE SEAT | THE SHAPE IS | POINT | POSITION | SIZE | WIDHT | LENGHT | HEIGHT | DEPTH | RADIUS | COLOR | MATERIAL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Hypothesized | 85,47% | 64,04% | 75,44% | 73,47% | 65,79% | 72,22% | 80,35% | 90,80% | 47,02% | 48,51% | 76,86% | 77,91% | 56,67% | 78,48% | 58,33% |
| Rejected | 4,27% | 8,77% | 23,68% | 23,47% | 23,68% | 27,78% | 8,09% | 4,31% | 33,77% | 41,04% | 17,36% | 19,77% | 40,00% | 17,72% | 12,50% |
| TEXT recognize | 10,26% | 27,19% | 0,88% | 3,06% | 10,53% | 0,00% | 11,56% | 4,89% | 19,21% | 10,45% | 5,79% | 2,33% | 3,33% | 3,80% | 29,17% |




| | COPY | DUPLICATE | DELETE | REMOVE | DIRECTION | DISPLACEMENT | MOVE | ORIENTATION | ROTATE | SCALE | ZOOM | OBJECT | HOME | CAMERA | CANCEL | UNDO |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Hypothesized | 59,65% | 50,00% | 71,15% | 64,29% | 66,67% | 50,00% | 74,14% | 70,00% | 86,67% | 96,88% | 64,88% | 82,76% | 66,15% | 50,00% | 60,00% | 75,00% |
| Rejected | 7,02% | 50,00% | 3,85% | 7,14% | 33,33% | 50,00% | 13,69% | 20,00% | 0,00% | 0,00% | 16,67% | 13,79% | 21,54% | 0,00% | 0,00% | 0,00% |
| TEXT recognize | 33,33% | 0,00% | 25,00% | 28,57% | 0,00% | 0,00% | 12,17% | 10,00% | 13,33% | 3,13% | 18,45% | 3,45% | 12,31% | 50,00% | 40,00% | 25,00% |

**Figure 5: Up: detected sentences for drawing; down: detected sentences for manipulation**

assistance during the drawing sessions. They felt frustrated, when the system did not respond, when they wanted to give a specific position or size.

## 5. CONCLUSION

In this paper, a multi-modal system has been presented that utilizes speech and gesture inputs to draw a 3D object in AutoCAD, using a Leap Motion and a microphone. After experimentation and evaluation, we found that speech and gestures are well-coordinated in human to human communication. Unfortunately, it is not the case with the devices that are used to interact with computers in MMIS. Our results indicate that performing a task using speech is perceived exhausting, when there is no shared vocabulary between a human and a machine, and the usability of traditional input devices supersedes the one of speech and gestures. Only a small ratio of participants, less than 7% in our experiments were able to carry out the tasks with appropriate precision.

Drawing with precision in a modeling software is more complex than expected. The speech recognition process is exhausting, when the system works slowly and does not respond properly. Speech recognition requires a simple grammar and a quiet environment to reduce the noise. Gestures seem to be more natural and less tiring to use in human-computer communication, if the users are able to use both hands, instead of one hand only. The system has to offer several gestures for the same action in order to satisfy most users. Even though the system was functional, we still noticed that it has sometimes lost track of gestures. People would prefer more natural interaction such as gesture and speech, if the performance of the equipment for the interaction could satisfy a standard level of operation. However, integration of a Leap Motion, speech, and AutoCAD did not match this standard.

## 6. REFERENCES

[1] AutoCAD. Objectarx developer's guide. *AutoDesk Inc*, 2002.

[2] R. A. Bolt. *"Put-that-there": Voice and gesture at the graphics interface*, volume 14. ACM, 1980.

[3] A. Erol, G. Bebis, M. Nicolescu, R. D. Boyle, and X. Twombly. Vision-based hand pose estimation: A review. *Computer Vision and Image Understanding*, 108(1):52–73, 2007.

[4] M. Kavakli and K. Nasser. Impacts of culture on gesture based interfaces. *International Journal on Advances in Life Sciences Volume 4, Number 3 & 4, 2012*, 2012.

[5] J. Liu and M. Kavakli. A survey of speech-hand gesture recognition for the development of multimodal interfaces in computer games. In *Multimedia and Expo (ICME), 2010 IEEE International Conf. on*, pages 1564–1569. IEEE, 2010.

[6] G. Marin, F. Dominio, and P. Zanuttigh. Hand gesture recognition with leap motion and kinect devices. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 1565–1569. IEEE, 2014.

[7] S. Oviatt. Multimodal interactive maps: Designing for human performance. *Human-computer interaction*, 12(1):93–129, 1997.

[8] S. Oviatt. Advances in robust multimodal interface design. *IEEE Comp. Graphics and Applications*, 23(5):62–68, 2003.

[9] J. A. Rodger and P. C. Pendharkar. A field study of the impact of gender and user's technical experience on the performance of voice-activated medical tracking application. *International Journal of Human-Computer Studies*, 60(5):529–544, 2004.

[10] F. Weichert, D. Bachmann, B. Rudak, and D. Fisseler. Analysis of the accuracy and robustness of the leap motion controller. *Sensors*, 13(5):6380–6393, 2013.